Published in final edited form as:

Autism Res. 2019 March ; 12(3): 495–504. doi:10.1002/aur.2068.

The Stability of Joint Engagement States in Infant Siblings of Children with and without ASD: Implications for Measurement Practices

Kristen Bottema-Beutel*,

Lynch School of Education, Boston College, Chestnut Hill, MA USA

So Yoon Kim,

Lynch School of Education, Boston College, Chestnut Hill, MA USA

Shannon Crowley,

Lynch School of Education, Boston College, Chestnut Hill, MA USA

Ashley Augustine,

Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN USA

Bahar Kecili-Kaysili,

Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN USA

Jacob Feldman, and

Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN USA

Tiffany Woynaroski

Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN USA

Abstract

Obtaining stable estimates of caregiver-child joint engagement states is of interest for researchers who study development and early intervention in young children with autism spectrum disorder (ASD). However, studies to date have offered little guidance on the numbers of sessions and coders necessary to obtain sufficiently stable estimates of these constructs. We used procedures derived from G theory to carry out a generalizability study, in which we partitioned error variance between two facets of our system for measuring joint engagement states: session and coder. A decision study was then conducted to determine the number of sessions and coders required to obtain g coefficients of 0.80, an a priori threshold set for acceptable stability. This process was conducted separately for 10 infant siblings of children with ASD (Sibs-ASD) and 10 infants whose older sibling did not have ASD (Sibs-TD), and for two different joint engagement states; lower- and higher- order supported joint engagement (LSJE and HSJE, respectively). Results indicated that, in the Sibs-ASD group, four sessions and one coder was required to obtain acceptably stable estimates for HSJE; only one session and one coder were required for LSJE. In the Sibs-TD group,

*Corresponding author Kristen.bottema-beutel@bc.edu.

Conflicts of Interest

The authors declare that we have no conflicts of interest

two sessions and one coder were required for HSJE; seven sessions and two coders were required for LSJE. Implications for measurement in future research are discussed

Lay Summary

This study offers guidance for researchers who measure joint engagement between caregivers and infants who have an older sibling with ASD, and who have older siblings who are TD.

Keywords

Autism spectrum disorder; G theory; joint engagement; stability; infant siblings

In early development, caregivers scaffold children's mutual attention to objects during play, as well as support children in coordinating their activities within joint play routines (Bakeman & Adamson, 1984; Bruner, 1982). During these joint engagement episodes, caregivers also introduce children to symbols, through both pretense and language (Bruner, 1975; Tomasello, 1995). Previous research has demonstrated that caregiver-child joint engagement offers an important context for the achievement of developmental milestones across various domains, in both typically developing (TD) children and children with autism spectrum disorder (ASD; Adamson, Bakeman, Deckner, & Ronski, 2009; Bakeman & Adamson, 1984; Bottema-Beutel, Yoder, Hochman, & Watson, 2014). Because of the purported influence of joint engagement on development, early intervention researchers have designed interventions that seek to optimize joint engagement between young children with ASD and interventionists or caregivers (e.g., Kaale, Smith, & Sponheim, 2012; Kasari, Freeman, & Paparella, 2006; Schreibman et al., 2015). These interventions are designed with the assumption that joint engagement will have a cascading effect on later developmental achievements, such as language, even after the intervention has stopped (Nordahl-Hansen, Fletcher-Watson, McConachie, & Kaale, 2016; Kasari, Gulsrud, Paparella, Hellemann, & Berry, 2015).

Recently, intervention research of this type has been extended to infant siblings of children with ASD, who are more likely to develop social communication and language impairments (Sibs-ASD; Ozonoff et al., 2011; Messinger et al., 2013; Rogers et al., 2014). The rationale for intervening on Sibs-ASD is that earlier intervention, provided prior to the time when ASD can be reliably diagnosed, may be more efficacious in supporting development than interventions that are provided later (Bradshaw, Steiner, Gengoux, & Koegel, 2015). In fact, planned work in our own laboratories will test the effects of an early intervention targeting foundational sensory function on higher level social communication and language, which may be mediated by caregiver-child joint engagement.

Reliably estimating the occurrence of joint engagement episodes that are most influential for development is therefore of interest for developmental and early intervention researchers alike. However, there is currently little research evaluating the *stability* of estimates of caregiver-child joint engagement that are derived from the semi-structured measurement contexts traditionally used by researchers (e.g., Parent-Child Free Play Procedure; PCFP; Bottema-Beutel et al., 2014). We use the term *stability* to mean the extent to which estimates

are associated with acceptable levels of measurement error (Cunningham, Preacher, & Banaji, 2001). A caveat of unstable measurements is that they can lead to increased Type II error, which occurs when the null hypothesis is not rejected when it is false (Cohen, Cohen, West, & Aiken, 2003). This can occur because, from a mathematical standpoint, the stability of an estimate for constructs such as joint engagement places an upper bound on its validity for detecting effects of interest (Crocker & Algina, 1986). Thus, ascertaining the stability of commonly used indices of caregiver-child engagement is critical for the design of research focused on this construct.

According to G theory, features of measurement systems can introduce variance that is not part of the construct of interest (i.e., error; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Different features of the measurement system that introduce error that can be assessed and potentially controlled are referred to as *facets* (Shavelson & Webb, 1991). Two of the most relevant facets for observational measurement systems, such as those used in the measurement of caregiver-child joint engagement, are coder and session (see Sandbank & Yoder, 2014 for a discussion of sampling contexts as an additional facet of measurement). In instances where stability is found to be unacceptably low for a single coder and observation session (i.e., what is commonly used in most research studies), scores for a given observational construct can be computed by multiple coders and taken from multiple observation sessions and then averaged together in order to increase the stability of the estimate.

Determining the Stability of Estimates Derived from Observational Measurement Systems

A *g* coefficient is a particular type of intra-class correlation coefficient that indexes the stability of a given estimate. A generalizability study can be conducted, which involves computing *g* coefficients from variance components generated from within-subjects ANOVA for each measurement facet (e.g., variance due to coder, variance due and session), the object of measurement (e.g., variance due to person), and all possible interactions. In general, *g* coefficients of 0.80 or higher are thought to reflect a level of stability that is sufficiently high to project against Type II error (Yoder, Lloyd, & Symons, 2018). In order to determine the number of coders and sessions from which to average scores to achieve an adequate *g* coefficient, decision studies, which use variance components to project *g* coefficients for different numbers of sessions and coders, can be conducted. Importantly, decision studies allow for inferring the stability that would be achieved with numbers of coders and sessions beyond that which is used within the generalizability study itself. For example, if two coders and two sessions are used to generate *g* coefficients, projections based on these variance components can be made for any number of sessions and coders (Shavelson & Webb, 1991; Yoder, Lloyd, & Symons, 2018).

It is widely understood that coders can introduce unwanted error into observational measurements. This type of error is usually dealt with by calculating inter-coder agreement statistics. A common statistic for variables using interval or ratio metrics is the intra-class correlation coefficient (ICC), usually computed using two-way random effects for single

coders (Hallgren, 2012). ICCs are computed on a subset of coded data, generally around 15–20% of the total sample, that has been coded by two independent coders. However, this procedure does not involve averaging scores from both coders in order to increase the stability of the estimate. Rather, inter-coder agreement statistics are instead computed to document that they are above a particular threshold, such as .80 suggested for ICCs ‘very good’ ICCs (Yoder, Lloyd, & Symons, 2018). Following these procedures, construct estimates from a single coder are then used in the analyses of interest. In contrast, decision studies allow researchers to determine the number of coders who should code *all* observational sessions for the purpose of producing an averaged, and more stable, estimate of each participant’s score.

Session is another possible source of measurement error, which is variability that is due to different aspects of the measurement context on a given day or observation, and not due to the inherent capacities of the participants being measured. Contexts that are highly unstructured (e.g., classrooms, home or community settings, laboratory assessments with few guidelines for participant/examiner behavior and/or other controls) may be considered more ecologically valid than highly structured lab contexts. However, the tradeoff is that unstructured contexts can introduce a large amount of error due to session (i.e., day-to-day variation in behavior); sometimes so much so that reliably measuring interactional constructs may not be feasible (Bottema-Beutel et al., 2014; McWilliam & Ware, 1994). The measurement procedures that are commonly used in the joint engagement literature (for example the *Brief Observation of Social Communication Change* [Nordahl-Hansen et al., 2016] and the *Communication Play Protocol* [Adamson, Bakeman, & Deckner, 2004]) do have some structure in that they involve standardized toy sets, lab play rooms, and instructions. However, they are relatively less structured than more heavily scripted assessments performed by highly trained members of a research team, such the *Autism Diagnostic Observation Schedule* as administered by research-reliable examiners (ADOS; Lord et al., 2012). Joint engagement states have the added complexity of being a dyadic construct, which means that there are two individuals (instead of one) who can perform differently across different sessions. It is therefore highly likely that measuring joint engagement within semi-structured free-play sessions entails measurement error due to session that researchers may wish to protect against.

Implications of Measurement Stability for Joint Engagement Research

In our past work, we have measured joint engagement states in order to test associations between different types of joint engagement and later developmental outcomes such as language. Of special interest to us has been two different types of *supported joint engagement*, which occurs when a child and caregiver play together with toys and the adult influences the child’s play, but the child does not reference the adult with gaze. Prior research has indicated that supported joint engagement that includes symbolic input (i.e., caregiver talk) is more influential for later language abilities than developmentally more advanced states such as coordinated joint engagement, which involves the child managing the interaction with gaze to the interaction partner’s face (Adamson et al., 2009; Bakeman & Adamson, 1984).

We have distinguished between two types of this particular joint engagement state; *higher-* and *lower- order supported joint engagement* to determine if we could further improve the specificity of the types of joint engagement that are most influential to development. In higher order supported joint engagement (HSJE), an adult and child reciprocally play together with toys, through observable processes such as turn-taking, imitation, or collaborating to achieve a commonly held goal. In lower order supported joint engagement (LSJE), the adult influences the child's play, but there is no reciprocity from the child. Our research (Bottema-Beutel et al., 2014) has indicated that HSJE is superior to LSJE in predicting expressive language. When these engagement states co- occur with caregiver talk about the child's focus of attention, HSJE is also superior to LSJE in predicting later receptive language. Recently, we found that when caregivers use verbs to talk about the child's focus of attention within HSJE, this supports children's later expressive verb use better than when caregivers use verbs to talk about the child's focus of attention in other engagement states (Crandall, Bottema-Beutel, McDaniel, Watson, & Yoder, under review).

We have also compared associations between HSJE and other constructs for groups of children with ASD and matched groups of TD children. We found that HSJE that co-occurs with caregiver talk about the child's focus of attention mediates the pathway between early expressive vocabulary and later receptive vocabulary, but only in children with ASD (Bottema-Beutel et al., 2018a). This finding indicates that children with ASD are especially reliant on the reciprocal engagement characteristic of HSJE in order for expressive language to influence later receptive language. We also found that sequential associations (the momentary influence of one event on another event) between HSJE and caregiver talk are greater for caregivers and children with ASD than for caregivers and TD children. This finding indicates that caregivers of children with ASD are especially careful to time their talk so that it occurs after instances when their child is particularly engaged with the caregivers in the context of toy play. On the other hand, caregivers of TD children may not be compelled to be especially mindful of when they talk about their children's play.

Differences in the stability of the two constructs that are being compared (e.g., HSJE versus LSJE, or HSJE in children with ASD versus HSJE in children who are TD) could be an alternative explanation for many of the findings that we have previously reported. For example, Bottema-Beutel et al. (2014) indicated that HSJE was a superior predictor of expressive language as compared to LSJE. If it were the case, however, that our estimate of LSJE was less stable than our estimate of HSJE, we may need to reconsider our interpretation of our past findings. Under these circumstances, it may not be the superiority of HSJE over LSJE as a developmental construct per se, but rather the relatively greater stability of our index of HSJE as compared to LSJE that led to our observation of greater associations with later language. A similar psychometric phenomenon could offer an alternative explanation for our past findings for between-group differences; if HSJE was less stable in TD children as compared to children with ASD, associations could be lower in TD children due to the relative instability of the HSJE estimate in TD children, and not because of a "true" difference in the magnitude of associations between groups.

Our previous studies were secondary data analyses in which only a single measurement session was conducted for each participant. Therefore, we could not partition variance in

scores due to session, which is needed in order to compute g coefficients with session as a facet of interest. Indeed, stability studies can be costly to conduct. It is likely that many researchers who are interested in deriving stable estimates of constructs like joint engagement do not have sufficient resources to collect multiple observation sessions for each participant, and to then employ and train multiple coders to code each session. Therefore, it can be helpful to make findings from stability studies publicly available, to aid other researchers in adequately planning data collection and analysis procedures in their own research.

In regards to our previous studies, we expected that the more developmentally advanced construct (HSJE) would be less stable than the less developmentally advanced construct (LSJE) in infants and children in the early stages of development, as there is some evidence that estimates become more stable as developmental achievements become more entrenched or consolidated (Sandbank & Yoder, 2014). We also expected that estimates derived from children with ASD or from children who are more likely to be diagnosed with ASD are likely less stable than estimates derived from TD children for similar reasons (i.e., because joint engagement would be expected to be less stable in children who are developmentally younger). This hypothesis would lead us to expect that our previous analyses of group differences would be more likely to be vulnerable to Type II error (failing to detect a superior association between HSJE and later language in children with ASD as compared to TD) than type I error. However, we test these assertions in the current study in order to increase confidence in our previous findings and to provide guidance for future research focused on joint engagement.

The Current Study

Studies of Sibs-ASD are beginning to give insight into the early developmental processes that precede diagnosis. Many of these children go on to receive a diagnosis of ASD, and even those who do not receive a diagnosis display characteristics consistent with the broader autism phenotype as early as 9 months of age (Elsabbagh et al., 2009; Ozonoff et al., 2011; Messinger et al., 2013). Given that constructs can be especially unstable when they are first emerging, measuring joint engagement could be especially subject to unwanted error at this very early stage of development. Estimating construct stability in order to optimize the measurement system is then especially important for Sibs-ASD. In the present study, we compared the stability of estimates of HSJE and LSJE between two groups of infants; those who had an older sibling with ASD (Sibs-ASD), and those who had an older sibling that did not have a diagnosis of ASD (Sibs-TD). We formed predictions about the stability of each construct within and between groups with similar logic to that described above. Specifically, we expected that HSJE would be less stable than LSJE, because HSJE is the more developmentally advanced state and less likely to be developmentally consolidated. Second, we expected that both constructs would be more stable in Sibs-TD as compared to Sibs-ASD for similar reasons. Even though groups were matched on mental age, we expected that given the social nature of joint engagement, these states would be less well-established in the sibling group as compared to the non-sibling group.

Method

Study Design

This study explored two facets of the measurement system that could be potential sources of error: session and coder. To examine these facets, a fully crossed design was used wherein each participant contributed two sessions of data, which was then coded for engagement states by two independent, blinded coders.

Participants

Participants were recruited from a large university in the southern United States, as part of an ongoing longitudinal correlational study on infant/toddler siblings of children with ASD. Institutional review board approval was granted prior to subject recruitment, and parental consent was obtained for all participant prior to data collection. From the larger study, twenty young children between the ages of 7 and 17 months were selected for participation, 10 in the Sibs-ASD group and 10 in the Sibs-TD group. Participants were selected to be matched on gender and mental age, with equal numbers of males ($n=5$) and females ($n=5$) in each group. Groups did not significantly differ on chronological age, mental age, expressive vocabulary, or receptive vocabulary (Wilcoxon rank sum test p 's $> .40$; see Table 1 for further detail regarding the demographic characteristics of each group).

Measures

Autism Diagnostic Observation Schedule (ADOS-2).—The ADOS-2 (Lord et al., 2012) was administered by a research-reliable examiner on the research team to confirm a diagnosis of ASD for older siblings of infants in the Sibs-ASD group (or confirmed via record review, in instances wherein a research reliable administration for the older sibling had been completed at the study site within the 12 months immediately preceding the infant's entry to the study).

Mullen Scales of Early Learning (MSEL).—The MSEL (Mullen, 1995), a standardized, norm-referenced, examiner led assessment of development intended for children birth-68 months of age, was used for the purposes of matching Sibs-ASD and Sibs-TD infants on mental age and more broadly characterizing the sample. The average age equivalency from four scales (i.e., fine motor, visual reception, expressive language, and receptive language) was used as the metric of mental age.

MacArthur-Bates Communicative Development Inventories (MCDI).—The MCDI (Fenson et al., 2003) was used to characterize early receptive and expressive language levels in our sample. The MCDI is a norm-referenced checklist of words commonly understood and used by young speakers of American English. In the Words and Gestures questionnaire, caregivers indicate words their child “understands” and words their child “says and understands”. The metric of receptive vocabulary was the number of words that the child was reported to understand (i.e., the sum of parent-reported words that the child “understands” and “understands and says”), and the metric of expressive vocabulary was the number of words that the child was reported to “say and understand”.

Parent-Child Free Play (PCFP).—HSJE and LSJE were coded from the PCFP, a 15 min, semi-structured caregiver-child interaction session. Participants were invited to play in a 10 foot by 10 foot room that included a clear plastic bin containing a standard set of toys. Parents were instructed to play with their child as they normally do at home using whichever of the toys that they wished, while taking care not to play with their backs to the camera. The bin including a variety of toys such as a doll, a toy barn and toy animals, and beaded necklaces. The PCFP was conducted twice for each participant, within a two week period.

Coding Procedures

To assess our joint engagement states measurement system, we used the standardized coding procedures that we have employed in prior studies of joint engagement states (Bottema-Beutel et al., 2014; 2018a; Bottema-Beutel, Lloyd, Watson, & Yoder, 2018b). Coders were two doctoral-level students who both had previous experience with individuals with ASD, although neither were experienced with very young children. Both were specializing in ASD-related topics in their doctoral work, and both were focusing on some aspect of social interaction as part of their work. Neither coder had been trained to use this coding system prior to the current study.

In the previous studies, both coders coded 20% of the study sessions for the purpose of computing inter-coder agreement; for this study, both coders coded 100% of the study sessions. Prior to coding the dataset for this study, coders were trained to be reliable with a previous dataset of PCFP video recordings, and were required to code with 80% accuracy (the ratio of the master file's duration and the trainee's duration multiplied by 100) on three consecutive files before coding videos collected for the present study. Video files were then divided into sets of five, and the two coders coded each file in the set. We chose sets of five to be consistent with our standard coding procedures, which allows for a reliability coder to code 20% of the total files if they randomly select one file from each set. Following the completion of each set, the first author randomly selected a file from the set, and determined if the two coders were within 80% agreement on duration for the two engagement states of interest. If the coders fell below this threshold, a discrepancy discussion was held to determine where the coders disagreed, and to resolve the issues apparently contributing to disagreement (see Yoder, Lloyd, & Symons, 2018 for a description of this process). The two coders then re-coded the remaining files in the set, and a new file was randomly selected to compute accuracy. This process continued until 80% agreement was reached for one file from the set, and was then repeated for each subsequent set. Coders were blind to group membership for all participants and to study hypotheses.

Analysis Procedures

Non-parametric statistics were used to compare group means on demographic characteristics (chronological age, mental age, receptive vocabulary, and expressive vocabulary) and the two engagement constructs of interest, duration of HSJE and LSJE. Non-parametric tests were chosen because of the small sample size, and the fact that data did not conform to the normality assumptions required of parametric tests.

ANOVA was used to compute sums of squares for each facet (person, session, and coder) and interactions between facets (person X session, person X coder, session X coder, and person X session X coder) separately for each group (Sibs-ASD and Sibs-TD), and separately for the two engagement states of interest (LSJE and HSJE). Although ANOVA is a parametric statistical procedure, our purpose was only to use the sums of squares generated by these procedures, not to use p values for the purpose of testing hypotheses. Stata statistical software was used for all computations (Stata Corp, 2017). This information was then imported into EduG version 6.1 (Swiss Society for Research in Education Working Group, 2012), a free online software, which was used to compute g coefficients for various numbers of sessions and coders. Facets were treated as random, to enable generalization of results to all possible sessions and coders. EduG computes both relative and absolute g coefficients; we report relative g coefficients, as we are most interested in the relative ordering of participants on each score and not the absolute value of the estimate. The raw data were also imputed into EduG to ensure reliability of the calculations; the results did not differ between these procedures.

Results

Preliminary Analyses

Wilcoxon rank-sum tests indicated that durations of HSJE and LSJE were not significantly different between groups (see Table 1).

Main Analyses

Sums of squares for each facet and for interactions between facets are listed in Table 2. Findings indicated that, for the Sibs-ASD group, HSJE was less stable than LSJE. Findings were reversed in the Sibs-TD group, with LSJE less stable than HSJE. Finally, HSJE was less stable in the Sibs-ASD group than the Sibs-TD group.

For the decision study, we restricted the number of coders to 1 or 2, and explored projected stability when varying the number of observations between 1 and 7 sessions. We did this because most of the variance was due to session and not to coder, which is consistent with prior studies of young children's engagement (McWilliam & Ware, 1994). Results indicated that in order to reach a g coefficient of 0.8 in the Sibs-ASD group, 4 sessions and 1 coder was required for HSJE, and 1 session and 1 coder was sufficient for LSJE (see Figure 1). In the Sibs-TD group, 2 sessions and 1 coder were required for HSJE, and 7 sessions and 2 coders were required for LSJE (see Figure 2).

Discussion

This study examined the stability of measurement systems for two engagement states of interest (HSJE and LSJE) collected for two populations of interest (Sibs-ASD and Sibs-TD). Using within-subjects ANOVA, we partitioned variance due to person, session, coder, and interactions between these facets. Decision studies were then conducted to determine the number of sessions and coders that would need to be averaged across in order to derive sufficiently stable estimates of each construct.

Several of the findings supported our predictions. LSJE appears to be more stable than HSJE, but only for the Sibs-ASD group. This is encouraging, as it suggests that conclusions drawn from studies of children with ASD that have found HSJE to be a superior predictor of language as compared to LSJE do not need to be revised based on the present findings regarding stability. It also confirms previous findings that for populations of children with ASD, developmentally earlier constructs tend to be more stable (Sandbank & Yoder, 2014).

LSJE in the Sibs-TD group may have been the least stable construct because, compared to HSJE, LSJE requires more “interactive work” from the caregiver to maintain the interaction relative to the interactive work required of the child. For example, if a child is playing with a set of beads, the caregiver can influence their play by jangling the beads above the child’s head so that the child reaches for them, or pulling and shaking the beads the child is holding. The child only has to accommodate this influence and does not need to extend any interactive overtures toward the caregiver. Therefore, the instability in LSJE Sibs-TD may be a result of the inconsistencies in *caregiver* interactive efforts rather than differences in the child’s proclivity for engagement from session to session. HSJE, on the other hand involves child feedback to parents (in the form of imitation or turn taking for example), which may have prompted caregivers of Sibs-TD to engage more consistently within this state. Because caregivers are less likely to have developmental concerns for children who do not have an older sibling with ASD, Sibs-TD caregivers may have sporadically extended efforts to engage their child when their child was not reciprocally engaging. This would explain the session-to-session variability in LSJE for Sibs-TD that we found, while HSJE had less session-to-session variability in this group.

In contrast, LSJE may have been quite stable in the siblings Sibs-ASD group because these caregivers may have consistently felt compelled to provide scaffolding for their child to jointly engage, whether or not the child was demonstrating reciprocity. Prior research has demonstrated that caregivers of children with ASD are more responsive to their children’s interactional needs in free play contexts than caregivers of children who are TD (Bottema-Beutel et al. 2018b; Bottema-Beutel, Malloy, Lloyd, Watson, & Yoder, 2018c). This responsiveness may extend to the consistency of their efforts to engage their child within LSJE.

HSJE was more stable in the Sibs-TD group than in the Sibs-ASD group, which also supported our prediction. Again, this is encouraging as it increases our confidence that previous group-difference findings showing superior associations between HSJE and other constructs of interest in children with ASD as compared to TD children were not likely due to measurement error. The relatively greater stability of HSJE in Sibs-TD relative to Sibs-ASD is possibly due to this engagement state being more established in the Sibs-TD group. It is of note that mean duration of HSJE was also slightly higher (although not significantly so) in the Sibs-TD as compared to the Sibs-ASD group, providing some evidence for our assertion.

In practical terms, if researchers would like to estimate one of the two engagement states we examined in a single group of infants for the purpose of detecting associations or intervention effects, they can select the number of sessions and coders that reach the $g =$

0.80 threshold in the corresponding panels of Figures 1 or 2. However, if they wish to compare two engagement states or compare engagement states between groups of children with ASD (or who have a higher likelihood of being diagnosed with) and TD children (or children with population-level likelihood of developing ASD), they should plan their measurement system so that it accommodates the least stable engagement state. Otherwise, researchers risk finding superior associations for the most stable engagement state simply due to relatively higher stability and not to truly superior associations.

Limitations

These results offer a starting point for researchers to plan adequate measurement systems, but there are at least three limitations that should be considered. First, it is not clear how these findings should be extended to research involving developmentally older participants. This is important to consider, as much early intervention research is focused on participants that are several years older than our infant samples. In general, observation sessions with older children may be associated with more stable estimates of joint engagement states, as developmental achievements are expected to become more consolidated over time (Sandbank & Yoder, 2014). We were somewhat surprised by the relative instability of LSJE as compared to HSJE in the Sibs-TD LR group, given that LSJE should be the developmentally more established state. We suggested that this difference may be because LSJE is highly dependent on the session-by-session inclinations of the caregivers, although this hypothesis will need to be further explored in future research. If it is indeed the case that caregivers are the primary contributors to the instability of LSJE in our Sibs-TD group, future research should also seek to determine how stability estimates change as children age in this group of children.

Second, it should be noted that the free play context we examined is somewhat less structured than other measurement systems that have been used to elicit joint engagement. For example, the *Communication Play Protocol* used by Adamson and her colleagues (e.g., Adamson et al., 2004) involve scripted interactions between caregivers and children over multiple “presses”. Introducing this kind of structure may increase the stability of estimates, which would necessitate fewer coders and sessions to reach a criterion g coefficient. In contrast, other joint engagement measurement systems have relatively less structure than the system examined here. For example, joint engagement measured in classroom (Boyd et al, 2018; Sparapani et al., 2016) or playground (Locke, Shih, Kretzmann, & Kasari, 2016) contexts would likely require additional sessions and coders than what we report here in order to achieve acceptable levels of stability (e.g., Bottema-Beutel et al., 2014; McWilliam & Ware, 1994).

Finally, this study used relatively small samples of only 10 participants per group. Small sample sizes can lead to imprecise estimates of G coefficients, but sample sizes similar to those used here (i.e., 10 per group) have been considered sufficient for estimating stability for similar observational measurement systems (e.g., Yoder, Woynaroski, & Camarata, 2016). Our results should be interpreted with this caution in mind.

Implications for Future Research

Our findings have relevance for researchers examining the developmental influence of joint engagement, and for researchers who measure joint engagement as a proximal intervention outcome (e.g., Kasari, Freeman, & Papparella, 2006) or as a mediator effect of treatment on more distal outcomes (Gulsrud, Helleman, Shire, & Kasari, 2016). We note that in our prior research, we were able to detect associations between HSJE and later language constructs with only a single coder and session. Indeed, the current study suggests that “true” effects may have been even larger than what we reported. If researchers wish to use more complex statistical procedures than what we have used (i.e., structural equation modeling to detect mediation effects within an RCT) to detect more nuanced relationships between HSJE and other variables, multiple measurement sessions to estimate engagement state durations may be required. In terms of future intervention research, we have previously suggested that researchers may wish to be more selective about the kinds of joint engagement formats that constitute their intervention outcomes, as there is some evidence that HSJE is a more developmentally valid construct than LSJE in that it bears stronger relationships with later language and social communication. However, at least in infant siblings, HSJE is a relatively less stable construct than LSJE. This may mean that if intervention researchers hone in on engagement states like HSJE, they may be required to include multiple measurement sessions in order to detect intervention effects on this outcome.

Conclusion

Our findings increase our confidence that previous demonstrations of the superior predictive value of HSJE relative to LSJE for later language and social communication was not likely due to issues of instability related to our measurement system/s. Likewise, results also provide evidence that previous group difference findings are not likely due to the relative instability of HSJE in TD children as compared to children with ASD. This study will aid researchers who are interested in deriving stable estimates of joint engagement states in planning for future studies by offering guidance on the number of sessions and coders to include in their measurement systems.

Acknowledgements

This work was supported by NIH U54 HD083211 (PI: Neul), CTSA award KL2TR000446 from the National Center for Advancing Translational Sciences (PI: Woynarowski) and NIH/NIDCD 1R21DC016144 (PI: Woynarowski).

References

- Adamson LB, Bakeman R, Deckner DF, & Ronski M (2009). Joint engagement and the emergence of language in children with autism and Down syndrome. *Journal of Autism and Developmental Disorders*, 39, 84–96. [PubMed: 18581223]
- Adamson LB, Bakeman R, & Deckner DF (2004). The development of symbol-infused joint engagement. *Child Development*, 75, 1171–1187. [PubMed: 15260871]
- Bakeman R, & Adamson LB (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development*, 55, 1278–1289. [PubMed: 6488956]
- Bottema-Beutel K, Lloyd B, Carter EW, & Asmus JM (2014). Generalizability and decision studies to inform observational and experimental research in classroom settings. *American Journal on Intellectual and Developmental Disabilities*, 119, 589–605. [PubMed: 25354126]

- Bottema-Beutel K, Lloyd B, Watson L, & Yoder P (2018b). Bidirectional influences of caregiver utterances and supported joint engagement in children with and without autism spectrum disorder. *Autism Research*. Advance online publication.
- Bottema-Beutel K, Malloy C, Lloyd BP, Louick R, Joffe-Nelson L, Watson LR, & Yoder PJ (2018c). Sequential associations between caregiver talk and child play in autism spectrum disorder and typical development. *Child Development*, 89, e157–e166. [PubMed: 28548711]
- Bottema-Beutel K, Woynaroski T, Louick R, Stringer Keefe E, Watson LR, & Yoder PJ (2018a). Longitudinal associations across vocabulary modalities in children with autism and typical development. *Autism*. Advance online publication.
- Bottema-Beutel K, Yoder PJ, Hochman JM, & Watson LR (2014). The role of supported joint engagement and parent utterances in language and social communication development in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 44, 2162–2174. [PubMed: 24658867]
- Boyd BA, Watson LR, Reszka SS, Sideris J, Alessandri M, Baranek GT, ... & Belardi K (2018). Efficacy of the ASAP intervention for preschoolers with ASD: A cluster randomized controlled trial. *Journal of Autism and Developmental Disorders*. Advance online publication.
- Bradshaw J, Steiner AM, Gengoux G, & Koegel LK (2015). Feasibility and effectiveness of very early intervention for infants at-risk for autism spectrum disorder: A systematic review. *Journal of Autism and Developmental Disorders*, 45, 778–794. [PubMed: 25218848]
- Bruner J (1975). The ontogenesis of speech acts. *Journal of Child Language*, 2, 1–19.
- Bruner J (1982). The organization of action and the nature of the adult-infant transaction In Tronick E (Ed.), *Social interchange in infancy: Affect, cognition, and communication*. Baltimore, MD: University Park Press.
- Cohen J, Cohen P, West SG, & Aiken LS (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Crocker L, & Algina J (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt.
- Cronbach L, Gleser G, Nanda H, & Rajaratnam M (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York, NY: Wiley.
- Cunningham WA, Preacher KJ, & Banaji MR (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, 12, 163–170. [PubMed: 11340927]
- Elsabbagh M, Volein A, Holmboe K, Tucker L, Csibra G, Baron-Cohen S, ... & Johnson MH (2009). Visual orienting in the early broader autism phenotype: Disengagement and facilitation. *Journal of Child Psychology and Psychiatry*, 50, 637–642. [PubMed: 19298466]
- Fenson L, Dale P, Reznick J, Thal D, Bates E, Hartung J, ... Reilly J (2003). *MacArthur communicative development inventories: User's guide and technical manual*. Baltimore, MD: Paul H. Brookes.
- Gulrud AC, Hellemann G, Shire S, & Kasari C (2016). Isolating active ingredients in a parent-mediated social communication intervention for toddlers with autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, 57, 606–613. [PubMed: 26525461]
- Kaale A, Smith L, & Sponheim E (2012). A randomized controlled trial of preschool-based joint attention intervention for children with autism. *Journal of Child Psychology and Psychiatry*, 53, 97–105. [PubMed: 21883204]
- Kasari C, Freeman S, & Paparella T (2006). Joint attention and symbolic play in young children with autism: A randomized controlled intervention study. *Journal of Child Psychology and Psychiatry*, 47, 611–620. [PubMed: 16712638]
- Kasari C, Gulrud A, Paparella T, Hellemann G, & Berry K (2015). Randomized comparative efficacy study of parent-mediated interventions for toddlers with autism. *Journal of Consulting and Clinical Psychology*, 83, 554–563. [PubMed: 25822242]
- Locke J, Shih W, Kretzmann M, & Kasari C (2016). Examining playground engagement between elementary school children with and without autism spectrum disorder. *Autism*, 20, 653–662. [PubMed: 26341991]

- Lord C, Rutter M, DiLavore P, Risi S, Gotham K, & Bishop SL (2012). Autism Diagnostic Observation Schedule, second edition (ADOS-2) manual (Part I): Modules 1–4. Los Angeles, CA: Western Psychological Services.
- McWilliam RA, & Ware WB (1994). The reliability of observations of young children's engagement: An application of generalizability theory. *Journal of Early Intervention*, 18, 34–47.
- Messinger D, Young GS, Ozonoff S, Dobkins K, Carter A, Zwaigenbaum L, ... & Hutman T (2013). Beyond autism: A baby siblings research consortium study of high-risk children at three years of age. *Journal of the American Academy of Child & Adolescent Psychiatry*, 52(3), 300–308. [PubMed: 23452686]
- Mullen EM (1995). Mullen Scales of Early Learning. Circle Pines, MN: American Guidance Service Inc.
- Nordahl-Hansen A, Fletcher-Watson S, McConachie H, & Kaale A (2016). Relations between specific and global outcome measures in a social-communication intervention for children with autism spectrum disorder. *Research in Autism Spectrum Disorders*, 29, 19–29.
- Ozonoff S, Young GS, Carter A, Messinger D, Yirmiya N, Zwaigenbaum L, ... & Hutman T (2011). Recurrence risk for autism spectrum disorders: A Baby Siblings Research Consortium study. *Pediatrics*, e1–e8.
- Rogers SJ, Vismara L, Wagner AL, McCormick C, Young G, & Ozonoff S (2014). Autism treatment in the first year of life: A pilot study of infant start, a parent-implemented intervention for symptomatic infants. *Journal of Autism and Developmental Disorders*, 44, 2981–2995. [PubMed: 25212413]
- Sandbank M, & Yoder P (2014). Measuring representative communication in young children with developmental delay. *Topics in Early Childhood Special Education*, 34, 133–141. [PubMed: 25364089]
- Shavelson RJ, & Webb NM (1991). Generalizability theory: A primer. Thousand Oaks, CA: Sage.
- Schreibman L, Dawson G, Stahmer AC, Landa R, Rogers SJ, McGee GG, ... & McNERNEY E (2015). Naturalistic developmental behavioral interventions: Empirically validated treatments for autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 45, 2411–2428. [PubMed: 25737021]
- Shavelson RJ, & Webb NM (1991). Generalizability theory: A primer. Thousand Oaks, CA: Sage.
- Sparapani N, Morgan L, Reinhardt VP, Schatschneider C, & Wetherby AM (2016). Evaluation of classroom active engagement in elementary students with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 46 782–796. [PubMed: 26433878]
- StataCorp (2017). Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC.
- Swiss Society for Research in Education Working Group. (2012). EduG (Version 6.1) [Computer software]. Retrieved from <https://www.irdp.ch/institut/english-program-1968.html>
- Tomasello M (1995). Joint attention as social cognition In Moore C and Dunham PJ (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Hillsdale, NJ: Erlbaum.
- Woynaroski T, Oller DK, Keceli-Kaysili B, Xu D, Richards JA, Gilkerson J, ... & Yoder P (2017). The stability and validity of automated vocal analysis in preverbal preschoolers with autism spectrum disorder. *Autism Research*, 10, 508–519. [PubMed: 27459107]
- Yoder PJ, Lloyd BP, & Symons F (2018). *Observational measurement of behavior* (2nd ed.). Baltimore, MD: Brookes Publishing.
- Yoder PJ, Oller DK, Richards JA, Gray S, & Gilkerson J (2013). Stability and validity of an automated measure of vocal development from day-long samples in children with and without autism spectrum disorder. *Autism Research*, 6, 103–107. [PubMed: 23436778]
- Yoder PJ, Woynaroski T, & Camarata S (2016). Measuring speech comprehensibility in students with down syndrome. *Journal of Speech, Language, and Hearing Research*, 59(3), 460–467.

Figure 1. Stability as quantified by g coefficients for different numbers of sessions and coders, for higher order supported joint engagement (HSJE; top panel) and lower order supported joint engagement (LSJE; bottom panel) in infants at greater likelihood (nearly 20 times the general population-level) of being diagnosed with autism spectrum disorder (infants with an older sibling diagnosed with autism spectrum disorder)

Figure 2. Stability as quantified by g coefficients for different numbers of sessions and coders, for higher order supported joint engagement (HSJE; top panel) and lower order supported joint engagement (LSJE; bottom panel) in infants at relatively low likelihood (general population-level) of being diagnosed with autism spectrum disorder (infants with typically developing older siblings).

Table 1:

Participant Characteristics by Group

	Sibs-TD			Sibs-ASD			<i>p</i> value from Wilcoxon rank-sum test
	<i>M</i>	<i>SD</i>	Range	<i>M</i>	<i>SD</i>	Range	
Chronological Age	10.7	2.41	7, 14	11.6	3.72	7, 17	0.68
Mental Age ^f	11.98	2.37	7.5, 14.75	11.15	2.65	7.25, 15	0.43
Expressive Vocabulary [‡]	5.7	6.65	0, 16	5.5	7.72	0, 20	0.91
Receptive Vocabulary [‡]	42.9	41.97	0, 110	60.8	64.77	3, 191	0.65
Duration of HSJE in sec	86.05	110.10	0, 346	52.78	42.01	0, 136	0.73
Duration of LSJE in sec	154.73	206.79	0, 479	136.20	96.75	13, 386	0.71

Note: Sibs-TD = participants with an older sibling who is typically-developing, Sibs-ASD = participants with an older sibling who has an ASD diagnosis, LSJE = Lower Order Supported Joint Engagement, HSJE = Higher Order Supported Joint Engagement

^f Mental age measured via Mullen Scales of Early Learning.

[‡] Vocabulary is raw number of words measured via MacArthur-Bates Communicative Development Inventories.

Table 2:

Partial Sums of Squares for Duration of HSJE and LSJE by Group

	Sibs-TD		Sibs-ASD	
	HSJE	LSJE	HSJE	LSJE
Person	402857.9 **	282703.7	68821.0	327841.4 ***
Session	52.9	18105.0	9.0	1232.1
Coder	739.6	9455.6	112.2	396.9
Person X Session	60190.1	105542.7	16315.2	26784.9
Person X Coder	5190.4	18977.1	688.0	2718.1
Session X Coder	608.4	950.6	0.63	336.4
Person X Session X Coder	3158.6	9013.1	464.6	5758.6

Note. Sibs-TD = infants with older siblings who are typically-developing, Sibs-ASD = infants with an older sibling who has a diagnosis of autism spectrum disorder, LSJE = Lower order supported joint engagement, HSJE = higher order supported joint engagement.

**
 $p < 0.01$,

 $p < 0.001$.